Et maintenant?
Utiliser des standards pour ne pas annoter en vain



# L'annotation : comment normaliser?

- formes de normalisation
- quelques acronymes du jour
- la TEI en particulier



# Combien de standards faut-il dans le monde?

| | |
|---|---|
| WKWBFY | un seul : solution centraliste |
| NWEUMP | aucun : solution anarchiste |
| FTH | autant qu'il en arrive : solution laissez-faire |



# Les normes ne s'imposent pas dans la vie intellectuelle

- soit ils emergent d'un besoin aperçu dans le communauté
- soit leur usage dérive du necessité d'utiliser une technologie particulière
- mais on ne renonce pas volontièrement à son indépendence!



# Standards : un paysage complexe

**Agences officiels de standardisation** nationales : AFNOR, ANSI, BSI, DIN; internationales: ISO, IEC, W3C, OASIS, TEI ...

**Regroupements des Personnes Interessées** Plusieurs... ex

- LISA (Localisation Industry Standards Association)
- MPEG (Moving Pictures Expert Group)

**Projets ayant des enjeux pre-normatifs** En Europe seul, on peut noter EAGLES, Multext, MATE, ISLE...

**Infrastructures de recherche** International: Bamboo, DARIAH, CLARIN; Français : TGIR-Corpus, Adonis



# Standards : on peut s'en passer?

Pour le scientifique, les standards pourraient sembler un inconvenience:

- ils figent les avances de la connaissance
- leur production est chronophage
- ... et nécessite des compétences sociales

quand même il y a des "plus" pratiques qu'il faut souligner:



# Quelques besoins scientifiques

1. Comment identifier et retrouver les ressources numériques d'interet linguistique sur le web?
2. Comment valider les résultats scientifiques obtenus par d'autres personnes?
3. Comment enrichir ou intégrer les ressources existantes avec ses propres idées?
4. Comment séparer les ressources des outils qui les gèrent/analisent?

Pour tout cela, les standards restent essentiels



# Normalisation des métadonnées (1)

Les livres, ça on connait. Pour les ressources linguistiques (corpus, lexiques, logiciels...) les règles de description sont toujours en train de s'établir.

- Les ressources linguistiques sont tres divers
- On peut garantir la pérennisation des bits -- mais pas forcément des structures/annotations qu'ils representent/contiennent
- Ce n'est encore pas evident *comment* décrire ces structures/annotation -- l'annotation resortissant nécessairement d'une théorie specifique



# Normalisation de métadonnées (2)

OLAC : http://www.language-archives.org/

- communauté dominé par linguistes de champs, des langues en danger, etc. : plusieurs ressources très variés et rarissimes.
- s'est donné également mission de gestion du communité
- se base sur d'autres standards (notamment Dublin Core et XML pour le contenu, et OAI-PMH pour sa distribution)
- à ne pas confondre avec ...



# IMDI

http://www.mpi.nl/IMDI/

- Effort pareil, ressortant des travaux du MPI de Nijmegen, plus centraliste, origine de plusieurs outils d'annotation et recherche sophistiqués, notamment pour les multimedia
- qui définit sa propre sous-ensemble des metadonnées
- se servant des mêmes technologies sousjacentes
- à ne pas confondre avec ...



# CLARIN

http://www.clarin.eu/




Une *infrastructure* (ou en anglais 'club') au niveau européen, profitant des existant activités standardisés, et visant de les améliorer.

# Mon papa est plus fort que le tien

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)

SITUATION:
THERE ARE
14 COMPETING
STANDARDS.







|< < PREV RANDOM NEXT > >|

PERMANENT LINK TO THIS COMIC: HTTP://XKCD.COM/927/
IMAGE URL (FOR HOTLINKING/EMBEDDING): HTTP://IMGS.XKCD.COM/COMICS/STANDARDS.PNG



# Standards are built on standards (1)

Standards de représentation fondamentale:

- W3C XML : représentation des arborescences (etc.)
- ISO 10646 : Unicode : représentation des caractères
- D'autres standards ISO définit les objets fondamentals:
  - noms de langage (ISO 639)
  - noms de pays (ISO 2166)
  - noms des systemes d'ecriture (ISO 15924)
  - représentation des dates (ISO 8601)
  - ...



# Standards are built on standards (2)

OAI-PMH
http://www.openarchives.org/OAI/openarchivesprotocol.html:

- définit un protocole pour l'exposé des métadonnées pertinentes a une ressource quelconque sur le web
- permet le moissonage des infos sur les ressources variées par les moteurs de recherche, et leur integration/affichage

Dublin Core : http://dublincore.org

- propose 15 champs d'infos minimales pour la description des ressources
- avec une syntaxe d'extension, qui s'expriment en XML

ISIDORE : bon exemple de moteur de recherche
http://www.rechercheisidore.fr



# Know your enemy

- En 1999 Bird et Liebermann ont fait le bilan de plusieurs formats d'annotation linguistique pour le Linguistic Data Consortium
- C'est un des premiers essais de définir une modèle abstraite de ce genre : la notion d'une *annotation graph* servant à représenter n'importe quelle annotation
- Plus recemment, sur http://annotation.exmaralda.org/index.php/Linguistic_Annotation on peut trouver une liste de presque une cinquantaine de formats toujours employés...
- La tour de Babel n'a pas encore disparu.



# Standardisation de contenu

Le problème: les objets encodés, ou annotés, sont très variables. Il n'y a que deux possibilités :

- A: on propose l'union de tous les possibilités
- B: on propose l'intersection de tous les possibilités

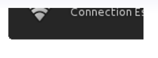
Le cas A nécessite une manière de simplifier/selectionner !

Le cas B nécessite une manière d'élargir/étendre !

Ou bien on adopte le principe de ‘format pivot’, en esperant de ne pas etre ‘perdu en traduction’



# Format de pivot



## Salt n' Pepper

(Romary & Zipser 2010)

Pepper (converter framework)

Salt (common metamodel)

Mapper

Mapper

Mapper

Mapp

Tiger XML

EXMARaLDA

PAULA XML

...

par exemple...



# Peut on definir une ontologie universelle linguistique?

Des essais...

- En 2003 Farrar et Langendoen ont proposé GOLD (General Ontology for Linguistic Data) : voir http://linguistics-ontology.org/
- ISO 24610 : structures de trait
- ISO 12620 : data category registry (specialisation de ISO 11179)
- W3C : Web Ontologie Language (OWL), et son expression en RDF

Mais il y a aussi la TEI...



# Les enjeux de la TEI

Mission quichotique : "Text Encoding for Interchange"

- faciliter la création, l'échange, et l'intégration des données textuelles informatisées
  - toute sorte de texte
  - toutes les langues
  - toute origine temporelle ou culturelle
- La TEI s'adresse également ...
  - aux débutants, cherchant des solutions bien connues et consensuelles
  - aux experts, cherchant à créer de nouvelles solutions



# Les non-enjeux de la TEI

D'origine, la TEI ne s'interessait pas à...

- le web (ca n'existait pas)
- la mise en page (tex, scribe...)
- l'intégration des pages-images/facsimilés numerisés
- la représentations des faits ou des objets (les bases de données)
- la production des logiciels

seul : les metadonnees, les textes, les analyses textuelles et linguistiques!



# Le paysage actuel de la TEI

- Structuration basique des textes continus
- Transcription diplomatique, images, multimédia, annotations...
- Données formelles : dates, noms de lieux ou de personnes...
- Données paratextuelles et "meta"
- Analyses linguistiques à tout niveau (y compris l'oral)
- Documentation de balisage
- Et cetera: voir `http://www.tei-c.org/P5/Guidelines/`

... un encyclopédie du balisage



# Un standard existe pour qu'on s'y conforme, non?

## The TEI Commandments

I. Thou shalt have no other encoding scheme but this one

II. Honour the consensus that thy days may be long in this land

III. Thou shalt not take the GIs of this scheme in vain

IV. Thou shalt not commit polysemy

*Text Encoding Initiative* 650 *November 1991*



# L'esprit TEI

Qu'est-ce que cela veut dire: «être conforme» à la TEI ?

- une pratique de balisage consensuel
- une lexique commune
- un respect de l'autonomie

La standardisation ne devait pas signifier «fais comme moi»; elle veut dire «expliques-moi ce que tu fais. »



# ... d'où les variations TEI

Par exemple: éléments pour description bibliographique: On a la choix entre

- <bibl> qui contient n'importe quel mélange de composants bibliographiques ... ou aucun
- <biblStruct> qui contient une sélection prédéfinie d'éléments, strictement structurée



# <bibl>: quelques exemplaires

```
<bibl>Blain, Clements and Grundy: Feminist
Companion to Literature in English (Yale, 1990)</bibl>
```

```
<bibl>
 <title>Dictionnaire des difficultés de la
   langue française</title> (<author>V. Thomas</author>,
<publisher>Larousse</publisher>)
</bibl>
```

```
<bibl>
 <idno type="ISO">ISO/IEC 2382 (all parts)</idno>,
<title type="introductory">Information technology</title>
 <title type="main">Vocabulary</title>
</bibl>
```



# <biblStruct>: un exemplaire

```xml
<biblStruct type="incollection">
 <analytic>
  <author>
   <forename>Pliny Earle</forename>
   <surname>Goddard</surname>
  </author>
  <title type="main" level="a">Athapascan (Hupa)</title>
 </analytic>
 <monogr>
  <editor>Boas, Franz</editor>
  <title> Handbook of American Indian Languages</title>
  <imprint>
   <pubPlace>Washington, D. C.</pubPlace>
   <publisher>Government Printing Office</publisher>
   <date>1911</date>
  </imprint>
  <biblScope type="pp">85-158</biblScope>
 </monogr>
</biblStruct>
```

(un des formats maintenant disponible de HAL)



# Les Guidelines augmentent le syntaxe avec une semantique

**vocabulaire** ses 521 éléments sont regroupés en 146 classes

**règles d'usage** ses 22 modules sont traduit en 7185 lignes de règles en Relax NG

**données contraintes** ses 21 datatypes et plusieurs règles formalisées en schematron

**règles de sélection** pas formalisées, mais documentées

**règles d'utilisation** *beaucoup* de prose

**règles maison** à construire intégralement.

> Comment gérer cette richesse?



# Il n'y a pas un seul “TEI dtd”

- TEI est un système *modulaire*. On s'en sert pour créer un système d'encodage selon ses propres besoins, en sélectionnant des *modules* spécifiques
- Chaque module définit un ensemble d'éléments (et leurs attributs)
- on peut sélectionner les éléments voulus, et même en changer des propriétés
- on peut y mélanger des éléments nouveaux, ou bien originels ou bien en provenance d'autres standards



# Architecture TEI

TEI ontologie

classe 34

classe 12

balises

modified element

module A

module B

schema

1



# Architecture TEI

- Le système TEI comprend plusieurs *modules*
- Chaque module comprend plusieurs *spécifications d'élément*
- Chaque spécification comprend:
  - un nom canonique (<gi>) en anglais pour l'élément et facultativememt des noms équivalents en d'autre langues
  - une description canonique de sa fonction ou mode d'emploi (facultativement traduite en d'autre langues) function
  - une déclaration pour chacune des *classes* auxquelles il appartient
  - une définition pour chacun de ses *attributss*
  - une définition de son *modele de contenu*
  - des exemples d'usage, des notes, des liens
- une spécification de *schéma* TEI (<schemaSpec>) se fait par une sélection de modules ou d'éléments, avec (eventuellement) des modifications
- un document TEI qui contient une spécification de schema s'appelle (informellement!) un *ODD* (One Document Does it all)



# Liste des modules

| Nom des module | chapitre |
|---|---|
| analysis | Simple Analytic Mechanisms |
| certainty | Certainty and Responsibility |
| core | Éléments Available in All TEI Documents |
| corpus | Language Corpora |
| dictionaries | Dictionaries |
| drama | Performance Texts |
| figures | Tables, Formulae, and Graphics |
| gaiji | Representation of Non-standard Characters and Glyphs |
| header | The TEI Header |
| iso-fs | Feature Structures |
| linking | Linking, Segmentation, and Alignment |
| msdescription | Manuscript Description |
| namesdates | Names, Dates, People, and Places |
| nets | Graphs, Networks, and Trees |
| spoken | Transcriptions of Speech |
| tagdocs | Documentation Éléments |
| tei | The TEI Infrastructure |
| textcrit | Critical Apparatus |
| textstructure | Default Text Structure |
| transcr | Representation of Primary Sources |
| verse | Verse |



# Comment choisir?

- On peut tout choisir! (pas vraiment une bonne idée)
- On peut partir d'une selection prédéfinie (TEI Lite, TEI Bare...)
- On peut faire artisanal -- selon les besoins spécifique de son projet

Dans ce dernier cas, il faut prendre conscience de toutes les possibilités disponibles...

Roma un logiciel en ligne qui simplifie cette procedure

http://www.tei-c.org/Roma/



# Pourquoi s'interesser toujours a la TEI?

Deux raisons pour lesquelles les standards échouent:

- ils sont basés sur une théorie pas encore mûre
- "not invented here": la communauté envisagée est trop diverse ou fragmentée



# Comment mûrir une théorie?

Dans son TEI ODD, on peut:

- limiter les valeurs possible d'un attribut plus ou moins strictement
- proposer des règles "schematron" sur le contenu (p.e. co-dependency)
- enlever quelques éléments facultatives
- ajouter de nouveaux éléments, labellisés dans votre propre espace de noms

Donc on peut évoluer et tester sa théorie, en restant toujours TEI-conforme.



# Addition d'éléments

Une schema ressemble à une grammaire, déjà existant. Comment serait-il possible d'ajouter des terminaux nouveaux?

- Les modeles de contenu s'expriment d'une manière indirecte
- Les définitions des éléments font référence normalement aux classes, et non pas directement aux éléments
- Donc pour ajouter un nouveau élément, on a besoin seulement d'identifier sa classification

La schéma TEI est enrichie de sémantique. Comment est-ce qu'on explique la signification d'un nouveau élément?

- une classe TEI peut indiquer une sémantique au dela de sa structuration
- la documentation ODD d'un element peut fournir des renseignements complètes sur sa signification.



# Not Invented Here?

- TEI P5 a des possibilités très extensives pour l'I18N...
- TEI héberge volontairement d'autres espaces de noms
- Donc on peut se servir des autres schémas existants:
  - SVG pour les graphiques
  - MathML pour le math
  - DCMI pour les metadonnees
  - ....
- La définition d'un élément TEI peut inclure (s'il y en a) sa mapping sur d'autres ontologies, formalisé par un élément <equiv> (equivalent)



# Organisation (sociale) de la TEI

TEI Consortium Members

TEI User Community

annual elections

proposals for change

Special Interest Groups

TEI Technical Council

TEI Executive Board

twice-yearly updates

TEI *Guidelines*



# L'évolution darwinienne, ça marche...

- faites vos modifications dans votre espace de noms
- documentez-les dans un ODD
- faites discuter vos propositions sur la liste TEI-L, ou dans un SIG!
- proposez les modifications efficaces au Conseil Scientifique de la TEI, en faisant un "feature request" sur sourceforge
- Il y a une version nouvelle de TEI P5 deux fois par an...



# Pour en savoir plus

- `http://www.tei-c.org`
- `http://tei.sf.net`
- `http://listserv.brown.edu/archives/cgi-bin/wa?SUBED1=tei-l&A=1`

plus, quelques références francophones:

- tei-fr@cru.fr
- http://meet.tge-adonis.fr
- http://lespetitescases.net/index102/
- http://www.culture.gouv.fr/culture/dglf/riofi/tei.htm
- http://artist.inist.fr/article.php3?id_article=122"/>

